Accepted for publication in the Journal of Biomedical Informatics
(Special issue on Auditing Terminologies).
(In press -- DOI:10.1016/j.jbi.2008.12.003).

Analyzing polysemous concepts from a clinical perspective: Application to auditing concept categorization in the UMLS

Fleur Mougin[1], *Olivier Bodenreider*[2$], *Anita Burgun*[3]

[1] LESIM, INSERM U593, ISPED, University of Bordeaux 2, France

[2] Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD, USA

[3] EA 3888, IFR 140, Faculté de Médecine, University of Rennes 1, France

[$]Corresponding author:

Dr. Olivier Bodenreider

National Library of Medicine

8600 Rockville Pike - MS 3841 (Bldg 38A, Rm B1N28U)

Bethesda, MD 20894 - USA

phone: 301 435-3246 - fax: 301 480-3035

olivier@nlm.nih.gov

Abstract

Objectives: Polysemy is a frequent issue in biomedical terminologies. In the Unified Medical Language System (UMLS), polysemous terms are either represented as several independent concepts, or clustered into a single, multiply-categorized concept. The objective of this study is to analyze polysemous concepts in the UMLS through their categorization and hierarchical relations for auditing purposes.

Methods: We used the association of a concept with multiple Semantic Groups (SGs) as a surrogate for polysemy. We first extracted multi-SG (MSG) concepts from the UMLS Metathesaurus and characterized them in terms of the combinations of SGs with which they are associated. We then clustered MSG concepts in order to identify major types of polysemy. We also analyzed the inheritance of SGs in MSG concepts. Finally, we manually reviewed the categorization of the MSG concepts for auditing purposes.

Results: The 1,208 MSG concepts in the Metathesaurus are associated with 30 distinct pairs of SGs. We created 75 semantically homogeneous clusters of MSG concepts, and 276 MSG concepts could not be clustered for lack of hierarchical relations. The clusters were characterized by the most frequent pairs of semantic types of their constituent MSG concepts. MSG concepts exhibit limited semantic compatibility with their parent and child concepts. A large majority of MSG concepts (92%) are adequately categorized. Examples of miscategorized concepts are presented.

Conclusion: This work is a systematic analysis and manual review of all concepts categorized by multiple SGs in the UMLS. The correctly-categorized MSG concepts do reflect polysemy in the UMLS Metathesaurus. The analysis of inheritance of SGs proved useful for auditing concept categorization in the UMLS.

Keywords

Biomedical terminologies, Auditing methods, Unified Medical Language System (UMLS), Polysemy, Semantic categorization.

1 Introduction

Ambiguity is a frequent issue in lexical representation, encountered in natural languages and reflected in terminological resources. Linguists make a distinction between contrastive and complementary ambiguity [1]. The former corresponds to homonymy, where "a lexical item accidentally carries several distinct and unrelated meanings". This is the case, for example, of the word *bank*, referring to both a financial institution and the sloping land beside a body of water. In contrast, complementary ambiguity "involves lexical senses which are manifestations of the same basic meaning of the word as it occurs in different contexts" [1]. Using the same example as earlier, the word *bank* also refers to both a financial institution (e.g., to open an account at the bank) and the building in which the banking business takes place (e.g., to go to the bank on First street). In this type of ambiguity, the various meanings of a lexical item are distinct, yet logically and sometimes systematically related. For this reason, it is often referred to as **polysemy** or systematic polysemy.

In lexical databases and terminologies organized by concept, ambiguity is reflected through the association of a given lexical item or term (i.e., an English word or word phrase) with several concepts. For example, in WordNet, the electronic lexical database of general English [2], the word *hospital* is ambiguous as it is associated with two concepts (called "synsets" in WordNet parlance, because they correspond to sets of synonymous lexical items). These two synsets are:

- *Hospital* (synonym: infirmary), defined as "a health facility where patients receive treatment", and whose ancestors (hypernyms in WordNet parlance) include *building edifice* and *physical object*; and
- *Hospital*, defined as "a medical institution where sick or injured people are given medical or surgical care", and whose ancestors include *institution*, *organization*, and *abstract entity*.

The top-level synsets in WordNet are called "unique beginners". The two distinct senses of *hospital* are linked to distinct unique beginners: *physical object* and *abstract entity*, respectively. Overall, some 13.5% of the nouns and almost half of the verbs are associated with more than one synset in version 3.0 of WordNet[1]. Similarly, in the Unified Medical Language System (UMLS) Metathesaurus, a terminology integration system for biomedicine [3], the word *cold* is ambiguous. This word is associated with seven distinct concepts[2], including *common cold*, a disease, *cold temperature*, a phenomenon, and *Chronic Obstructive Lung Disease*, whose acronym is "COLD". Overall, some 50,000 terms are associated with more than one concept in the UMLS.

The two forms of ambiguity, homonymy and polysemy, are represented in different ways in different lexical resources. One major difference pertains to the explicit representation of relations among the various concepts for polysemous lexical items, i.e., those lexical items related in more or less systematic and predictable ways. Another difference is whether ambiguity is consistently represented throughout a given resource. As shown earlier, WordNet does not explicitly distinguish between ambiguous and polysemous lexical items. In both cases, the corresponding lexical items are associated with several

[1] http://wordnet.princeton.edu/man/wnstats.7WN
[2] UMLS concepts are presented in *italic typeface*

distinct synsets, and no relation is recorded among synsets for polysemous lexical items. In contrast, in CoreLex, a derivation of WordNet, lexical items are grouped into systematic polysemy classes instead of being simply associated with distinct and unrelated synsets [4]. In other words, WordNet does not account for any regularity between the multiple senses of polysemous terms. Indeed, the two synsets presented above for the polysemous word *hospital* are fully independent.

In the UMLS, ambiguous terms are not always represented consistently. On the one hand, lexical items having distinct senses are expected to be represented in distinct concepts [5]. For example, the seven senses of *cold* presented above are associated with as many distinct concepts in the UMLS Metathesaurus. In most cases of homonymy and polysemy, lexical items are associated with distinct concepts. However, this is not systematically the case and polysemous lexical items sometimes belong to the same concept. Indeed, considering again the word *hospital*, there is only one concept *Hospitals* in the UMLS Metathesaurus for the two senses mentioned above. In this case, polysemy is indicated by the multiple categories (semantic types[3]) assigned to this concept. In fact, the single concept *Hospitals* is categorized both as a Manufactured Object and a Health Care Related Organization.

In summary, polysemous terms are not always represented consistently across terminological systems. For example, the polysemous word *hospital* is represented by two distinct synsets in WordNet, located in two different hierarchies, whereas it is represented by only one concept in the UMLS Metathesaurus, categorized by two distinct high-level categories. Differences in representation can also be observed within a given terminological system where polysemous terms can be represented by a single, multiply-categorized concept (e.g., *Hospitals* in the UMLS) or by several distinct concepts (e.g., the enzyme *glycosyltransferase* and the catalytic function it supports are represented by distinct concepts in the UMLS).

The objective of this study is to analyze polysemous concepts in the UMLS through their categorization and hierarchical relations. More precisely, we take advantage of the Semantic Groups for identifying polysemous concepts and study how the multiple semantic groups of polysemous concepts are inherited. We show that insights gained from studying polysemous concepts can be used for auditing purposes.

2 Background

2.1 UMLS

The Unified Medical Language System® (UMLS®) [6] includes two sources of semantic information: the Metathesaurus® and the Semantic Network. The UMLS Metathesaurus was assembled by integrating almost 150 source vocabularies. It contains about 1.5 million concepts, i.e., clusters of synonymous terms coming from multiple source vocabularies identified by a Concept Unique Identifier (CUI). More than 36 million relations are recorded between these concepts. Several types of relationships[4] among concepts are

[3] Semantic types are presented in sans serif typeface

[4] Relationships among concepts are presented in *italic, sans serif typeface*

recorded in the Metathesaurus: *parent* / *child of* (PAR / CHD) and *broader* / *narrower than* (RB / RN) essentially correspond to hierarchical relations, while the other relationships are associative. More than 7.5 million hierarchical relations are represented in the Metathesaurus.

The Semantic Network is a much smaller network of 135 Semantic Types (STs) organized in a tree structure [7]. Each Metathesaurus concept is assigned at least one ST. Groupings of STs, called Semantic Groups[5] (SGs), represent subdomains of biomedicine such as Anatomy, Chemicals & Drugs, and Disorders [8]. Each ST belongs to one and only one SG.

2.2 Multiple categorization and polysemy

2.2.1 Multiple semantic types

Some UMLS concepts are categorized by several STs. Justification for multiple categorization in the UMLS is threefold. First, a concept can be multiply-categorized *by convention* if one of its constituting terms is polysemous in the source vocabulary from which it comes. Indeed in the UMLS, STs are assigned to reflect the meaning of terms in their original source [9]. If a given term has multiple meanings in a source vocabulary, the corresponding concept is expected to be assigned multiple STs accordingly. For example, *Books, Illustrated* (C0006003) is categorized by Manufactured Object and Intellectual Product because it is defined in the Medical Subject Headings (MeSH) both as a book and an audiovisual aid. The assignment of multiple STs to *Books, Illustrated* thus reflects in the UMLS Metathesaurus the polysemy existing in its original source vocabulary.

Multiple categorization is sometimes the consequence of the *integration* achieved by the UMLS. To avoid creating concepts among which distinctions are not significant for clinical purposes, the UMLS developers sometimes lump under the same concept terms from several source vocabularies exhibiting only minor differences in their definition. In this case, the distinct meanings of each original term is represented by assigning these concepts different STs. (This case differs from multiple categorization *by convention* in that, here, the terms need not be polysemous in the source vocabularies). The concept *Loss of Heterozygosity* (C0524869) is thereby assigned to the STs Cell or Molecular Dysfunction and Genetic Function, reflecting both the "loss of one allele at a specific locus, caused by a deletion mutation; or loss of a chromosome from a chromosome pair, resulting in abnormal hemizygosity" (definition from MeSH) and the "genetic phenomenon due to deletion or mutation in one allele of a polymorphic gene, as detected by expression after cell fusion" (definition from the CRISP thesaurus), respectively. Because it integrates the terms from MeSH and CRISP, the UMLS concept *Loss of Heterozygosity* is polysemous, denoting both the disease caused by the deletion mutation and the genetic function.

Finally, multiple categorization can also be *intentional*. Some terms exhibit multiple semantic features (or facets) represented by multiple STs, often one sortal type (denoting its essence) and one or more role type (denoting its function). Instead of representing such terms as distinct concepts, the UMLS represents them as a unique concept categorized by multiple STs, each of them denoting some semantic feature of the term. In fact, almost all chemical concepts are assigned, by design, several STs for representing their

[5] Semantic groups are presented in bold, sans serif typeface

structural and functional features. For example, *progesterone* (C0033308) is categorized simultaneously as Steroid (structure), Hormone (function) and Pharmacologic Substance (function). Here, *progesterone* is not a polysemous concept but rather a concept having multiple facets.

The assignment of multiple STs to a concept may denote polysemy, but not systematically. Multiple categorization *by convention* and from *integration* generally corresponds to polysemy, while *intentional* multiple categorization does not. Multiple ST categorization is a frequent occurrence in the UMLS, where nearly 15% of the concepts are categorized by several STs. A large proportion of these concepts corresponds to chemicals (*intentional* multiple categorization).

2.2.2 Multiple semantic groups

The SGs have been designed to represent subdomains of biomedicine, by clustering the STs into 15 groups (Table 1). One of the principles underlying the creation of the SGs is **exclusivity** [10]. SGs are pairwise disjoint, i.e., each ST belongs to one and exactly one SG, and the SGs are expected to realize a partition of UMLS Metathesaurus concepts. Other principles include **semantic validity** and **completeness** [8]. The subdomains delineated by the SGs are thus expected to be semantically coherent and to provide exhaustive coverage. By grouping STs across hierarchies in the Semantic Network, SGs are designed to absorb a large part of the multiple ST categorization, especially the *intentional* multiple categorization.

The vast majority of these multiply-categorized concepts are associated with only one SG and only 0.08% of all UMLS concepts are associated with more than one SG. This is in contrast to the frequency of multiply-categorized concepts (15%) and shows that the STs assigned simultaneously to concepts tend to belong to the same SGs.

From the perspective of lexical semantics, the SGs already absorb some of the systematic polysemy reflected through the multiple categorization of concepts with STs. For example, *Cleft palate* (C0008925), the "Congenital fissure of the soft and/or hard palate, due to faulty fusion" is categorized by both Congenital Abnormality (because this birth defect results from abnormal fusion of tissues during fetal development) and Disease or Syndrome (because it can cause feeding difficulties and can be treated surgically). Despite this dual categorization, *Cleft palate* is associated with only one SG (Disorders), in which the STs Congenital Abnormality and Disease or Syndrome were purposely grouped.

Therefore, the association of a concept with several distinct subdomains of biomedicine provides a convenient way of identifying residual polysemous concepts (i.e., polysemy *by convention* and resulting from *integration*). In practice, polysemous concepts can be easily identified when they are categorized by STs from distinct subdomains, i.e., when they are associated with more than one SG. As shown in Figure 1, one such concept, *Medical center* (C0565990), is categorized by the STs Manufactured Object and Health Care Related Organization, associated with two distinct SGs Objects and Organizations, respectively. As mentioned earlier, this is a case of polysemy, not homonymy, because the two meanings of *Medical center* are related.

Our study focuses on those concepts associated with more than one SG. Such concepts are hereafter referred to as **multi-SG (MSG) concepts**. We hypothesize that these concepts exhibit some form of polysemy. Note that MSG concepts do not correspond to all polysemous concepts existing in the UMLS;

in particular, some concepts categorized by more than one ST are also polysemous. We however restrict this work to MSG concepts since they represent a smaller sample among which the proportion of polysemous concepts is expected to be larger than among multi-ST concepts in general.

2.3 Inheritance principles in biomedical terminologies

Hierarchical relations form the backbone of biomedical terminologies. The subsumption relationship (also called taxonomic relationship or *isa*) holds between a more specific concept and a more generic concept when all instances of the more specific concept are also instances of the more generic concept [11,12]. In lexical semantics, as illustrated by examples from WordNet presented earlier, the taxonomic relationship is called hypernymy [13]. From a classification perspective, the more specific concept has all the properties of the more generic concept plus some specific properties (differentiae) [14]. Such differentiae are often reflected in textual and formal definitions. More specifically, differentiae usually result from the introduction of a new property in the more specific concept or from the refinement in the more specific concept of one property of the more generic concept [15].

Categorization can be seen as one of the properties of biomedical concepts. According to the principle of inheritance of properties presented above, a more specific (child) concept is expected to be categorized like the more generic (parent) concept or to have a more specific category, and can exhibit differentiae. In terms of ST categorization in the UMLS, this implies that a concept whose parent is categorized by a given ST is expected to be categorized by the same ST or one of its descendants and possibly by one or more additional ST(s). One of the following three outcomes is expected when comparing the categorization of child and parent concepts in the UMLS Metathesaurus. The first case is when both concepts have the same categorization, such as *Viral meningitis* (C0025297) which inherits its ST, Disease or Syndrome, from *Meningitis* (C0025289). Secondly, the child is categorized by a descendant of the ST categorizing its parent. For example, *Tumors of Adrenal Cortex* (C0001618) is assigned Neoplastic Process which is a (direct) descendant of the ST Disease or Syndrome to which its parent *Adrenal Gland Diseases* (C0001621) is assigned. Finally, the child concept can be categorized by some ST(s) in addition to the ST(s) of its parent (or one of its descendant ST(s)). For example, *Butter* (C0006494) and its parent *Dairy Products* (C0010947) are both categorized by Food, and *Butter* is additionally categorized by Lipid.

Of course, the inheritance principles outlined above are only applicable to concepts among which the parent-child relation is a true subsumption (*isa*) relationship, which is not always the case in biomedical terminologies. Hierarchies are often created to support a specific purpose (e.g., information retrieval) and the relation between parent and child concepts cannot always be assumed to be *isa*. Moreover, as shown earlier, the editorial rules for multiple categorization are not homogeneous throughout the Metathesaurus. Therefore, the inheritance principles can be used, at best, as guidelines for interpreting the semantic consistency (or lack thereof) between the categorization of child and parent concepts.

2.4 Related work

Auditing semantic categorization in the UMLS has been the object of active research in the past decade. Several auditing procedures leverage multiply-categorized concepts and assess whether multiple categorization is valid or reflects miscategorization. The hierarchical organization of concepts in the

UMLS Metathesaurus is also compared to hierarchical relations between the corresponding STs in the Semantic Network in order to identify inconsistencies.

Based on ST combinations, Gu et al. proposed a model for representing the UMLS which enabled them to characterize different kinds of miscategorization [16]. Expanding on this work, they performed a large auditing in order to give precise statistical results about UMLS categorization [17]. In order to limit the need for manual review, they first constituted meta-STs which group STs hierarchically when they share structural and semantic characteristics [18]. Concepts associated with more than one meta-ST were reviewed by an expert. The authors showed that infrequent ST combinations were more likely to correspond to errors. Causes for multiple categorization included ambiguity, polysemy, missing and redundant categorizations, and miscategorization.

Cimino also proposed a method to audit the UMLS categorization by defining exclusivity between STs, based on their definition [19]. Some concepts categorized by several mutually exclusive STs exhibited categorization problems, such as ambiguity and inconsistency. In addition, the author analyzed UMLS categorization indirectly by comparing hierarchical relations between Metathesaurus concepts with those asserted between their corresponding STs. This approach, further investigated in [20], was useful to detect not only wrong and missing assignments, but also cases where the ST(s) of the parent concept was more specific than the ST(s) of the child concept. A few cases of missing hierarchical relations between STs in the Semantic Network were also identified.

In our study, we use the UMLS SGs rather than STs or hierarchical groupings of STs as the basis for assessing polysemy and auditing categorization. Our hypothesis is that the SGs were created to absorb part of the polysemy frequently encountered in the biomedical domain (e.g., between Disease or Syndrome and Congenital Abnormality). Our method is thus designed to focus on residual polysemy and errors. Like Cimino, we examine pairs of concepts in hierarchical relation and analyze the categorization of the child concept in relation to that of the parent concept. However, we concentrate our efforts on the inheritance of SGs.

3 Methods

In order to analyze the characteristics of polysemous concepts, we use the association of a concept with multiple SGs as a surrogate for polysemy. We first describe how we extract MSG concepts from the UMLS Metathesaurus. We then cluster MSG concepts in order to identify major types of polysemy. We audit the MSG concept categorization through inheritance in a quantitative way. Finally, we perform a qualitative auditing of the MSG concept categorization.

3.1 Identifying MSG concepts

The method we use to identify MSG concepts is straightforward. For each UMLS concept, we simply look up its ST(s) and select the concepts whose STs belong to distinct SGs. For example, the concept *Medical center* (C0565990), mentioned earlier, is identified as a MSG concept, because it is categorized by two STs from distinct SGs: Manufactured Object from the SG Objects and Health Care Related Organization from the SG Organizations (Figure 1).

3.2 Clustering MSG concepts

In order to characterize MSG concepts from a qualitative perspective, we cluster them according to hierarchical relations among them and in an attempt to establish a limited number of semantically homogeneous clusters.

In practice, because some Metathesaurus concepts are not hierarchically related to any other concepts, we start by identifying such concepts among the MSG concepts. In addition to concepts isolated in the Metathesaurus (i.e., those with no parent or child concepts), some of these concepts are isolated among the MSG concepts (i.e., those with no MSG parent or child concepts).

For each of the remaining MSG concepts, we create semantically homogeneous clusters (i.e., clusters in which all concepts are associated with the same combination of SGs), by exploring the *parent of* and *broader than* relationships recursively, adding to the cluster those ancestors sharing the same SG categorization as the source concept. The clusters so obtained are then merged in order to remove redundancy. Towards this end, **vertical** merging eliminates clusters that are completely nested in other clusters (Figure 2 (a)). Finally, in order to facilitate the analysis, **horizontal** merging of clusters aggregates two clusters – including singleton clusters – in the following circumstances: when two clusters share a common concept (Figure 2 (b) – bottom part) and when two concepts, one from each cluster, share a common parent (Figure 2 (b) – top part).

Semantic characterization of polysemous concepts. For each combination of SGs, clusters containing at least 10 concepts are examined in order to identify semantic patterns. In particular, we note whether one combination of STs is predominant throughout the clusters. (In practice, a pair of STs is deemed predominant for a series of clusters if at least 50% of the MSG concepts in these clusters are categorized by these STs.) When no predominant ST combination is found, the top ST combinations are listed instead.

3.3 Quantitative auditing of concept categorization through inheritance

We take advantage of the semantically homogeneous clusters built earlier to audit the inheritance of SGs in MSG concepts. Although MSG concepts violate the exclusivity principle of the SGs, they are expected to comply with general inheritance principles. In other words, while MSG concepts are subsumed by disjoint parents, they should inherit properties from their parents nonetheless. In particular, we want to compare the SG categorization of MSG concepts to that of their parents.

Applied to SG categorization of MSG concepts, the inheritance principle outlined in section 2.3 can be restated as follows.

Inheritance principle: The SG categorization of a MSG concept is inherited from its parent(s).

There are two corollaries to this principle, defining the properties of inheritance from the perspective of parent concepts and children concepts.

Corollary 1: The SG categorization of a MSG concept is inherited either from a unique parent or from multiple parents.

In practice, when the SG categorization of a given MSG concept is inherited from a unique parent, this parent must be a MSG concept of the same kind as the original MSG concept (i.e., associated with the identical SG combination). For example, as shown in Figure 3 (a), the UMLS concept *Periodicals as Topic* (C1956227) inherits its SG categorization from its unique parent *Serial Publications* (C0036719) and is thus associated with the same SGs, namely Concepts & Ideas and Objects.

In contrast, when the SG categorization of a given MSG concept is inherited from multiple parents, the MSG concept inherits each SG from different parents and its SG categorization results from the combination of its parents' SG categorization. For instance, as shown in Figure 3 (b), the concept *Potassium, Dietary* (C0162800) is a child of both *Potassium Compounds* (C0206140 and ST: Inorganic Chemical), which is associated with the SG Chemicals & Drugs, and *Dietary mineral* (C0596444 and ST: Food), associated with the SG Objects. The concept *Potassium, Dietary* inherits one ST from each parent and is therefore associated with both SGs Chemicals & Drugs and Objects.

Corollary 2: The SG categorization of a MSG concept is transmitted to its descendant(s).

Implied by this corollary is that every descendant of a MSG concept is expected to be associated with the same SGs as its parent concept. For example, revisiting Figure 3 (a) from the perspective of the child concept, the MSG concept *Serial Publications* (C0036719) is associated with the SGs Concepts & Ideas and Objects. Its child concept *Periodicals as Topic* (C1956227) is also associated with the same two SGs.

According to these inheritance principles, all the descendants of the root of a given cluster should be part of the cluster seeded by this root concept. In practice, in order to avoid redundancy in our analysis, we start from the roots of the large clusters (prior to merging clusters horizontally). For each root concept, we compute the list of all descendants, recursively, using both *child of* and *narrower than* relationships. From the set of all the descendants of a root concept associated with a pair of SGs (SG_1, SG_2), we compute the proportion of descendants associated with the same pair of SGs. Additionally, we compute the proportion of descendants associated with one of the root's SGs, i.e., SG_1 or SG_2, and the proportion of descendants associated with at least one SG different from those of the root concept.

3.4 Qualitative auditing of concept categorization and polysemy

In order to complement the quantitative analysis of MSG concepts, we performed a qualitative auditing in order to check if these concepts are indeed polysemous or miscategorized and, when applicable, to identify the origin of the polysemy. Toward this end, two authors (OB and AB), who are medical doctors, performed a manual analysis of the MSG concepts. In practice, each concept was audited independently by the two authors and a consensus was established through discussion when necessary. We examined the categorization of each concept in relation to that of its parent concepts in the original source vocabularies. More precisely, we established the validity of the categorization of the MSG concepts and assessed the inheritance of the categorization from their parent concepts (consistent, inconsistent, or incomplete). We considered polysemous those MSG concepts for which the multiple categorization is correct.

Additionally, we attempted to determine which editorial rule is at the origin of the polysemy: polysemy *by convention* or *due to integration* (see 2.2).

4 Results

4.1 Identifying MSG concepts

In the 2008AA version of the UMLS, of the 1,468,958 active Metathesaurus concepts, only 1,208 concepts (0.08%) are associated with more than one SG (MSG concepts), while 1,467,750 concepts (99.92%) are associated with one and only one SG. Table 2 shows the repartition of the 1,468,958 Metathesaurus concepts by SG.

In order to characterize the 1,208 MSG concepts, we first list the combinations of SGs with which they are associated. All MSG concepts are categorized by at most two STs and are consequently associated with at most two SGs. Moreover, among the 105 possible pairs of SGs resulting from the combination of 15 SGs, only 30 SG pairs were observed in practice. The number of MSG concepts for each combination of SGs is listed in Table 3. Four combinations of SGs account for 75% of all MSG concepts and the 95% of MSG concepts are covered by fifteen SG pairs.

As illustrated in Figure 4, the 15 most frequent pairs of SGs can be displayed in a network where nodes are SGs and relations correspond to the combination of SGs with which MSG concepts are associated. This network shows a small number of high degree SGs, especially Concepts & Ideas and Objects. These two SGs are hubs in the network (each of them is combined with many other SGs in the categorization of MSG concepts).

4.2 Clustering MSG concepts

Originally, there were 693 singleton clusters (concepts isolated in the UMLS or among the MSG concepts). Resulting from the clustering process are 276 singleton clusters (417 isolated concepts were merged horizontally) and 75 clusters of at least two concepts comprising the remaining 932 MSG concepts. Table 4 shows the number of singleton and non-singleton clusters for the 15 top pairs of SGs. The distribution of the number of concepts per cluster is shown in Figure 5. Overall, 78.6% of the clusters are singletons and 98% of the clusters contain at most 10 concepts. The largest cluster comprises 424 concepts and two other clusters contain more than 100 concepts.

Extended example. One of the clusters we obtained is displayed in Figure 6. This cluster comprises eight MSG concepts, all categorized by the STs Social Behavior from the SG Activities & Behaviors and Mental or Behavioral Dysfunction from the SG Disorders. One singleton cluster, which comprises the concept *Sexual harassment* (C0162790), was merged horizontally with the cluster C_1, rooted by the concept *Violence* (C0042693), because *Sexual harassment* and *Violence* share a common parent: *Aggressive behavior* (C0001807). It must be noted that this common parent contributes to the aggregation of two clusters of MSG concepts despite the fact that it is not itself a MSG concept. The former singleton cluster is removed from the final list of singleton clusters after its merger with cluster C_1. C_1, enriched with the former singleton cluster, and C_2, rooted by the concept *AODR interpersonal and societal problems* (C0683066),

share one concept, namely *AODR violence* (C0814667), child of both *AODR interpersonal and societal problems* and *Violence*. This shared concept forms the basis for merging the clusters C_1 and C_2 (horizontally). The final cluster results from the merger of three original clusters and comprises eight concepts, all associated with the two SGs Activities & Behaviors and Disorders.

Semantic characterization of polysemous concepts. Table 5 presents the predominant (or top) ST pair(s) for the clusters from the 15 SG pairs containing at least 10 concepts. For example, for the pair of SGs Objects-Organizations, 406 MSG concepts (86%) are categorized by both Manufactured Object and Health Care Related Organization. This predominant ST combination is thus deemed representative of the kind of polysemy identified in clusters of concepts associated with the pair Objects-Organizations. Conversely, no predominant ST combination is found for the pair Chemicals & Drugs-Objects, the top ST combinations are thus listed instead, namely Lipid-Food (in 34.5% of cases), followed by Pharmacologic Substance-Food and Hazardous or Poisonous Substance-Manufactured Object.

4.3 Quantitative auditing of concept categorization through inheritance

We analyzed in detail the 132[6] MSG concepts found as single roots of clusters. We obtained different categories for these MSG concepts: all of their descendants are all MSG concepts of the same kind (full compatibility); only some of their descendants share the same SG categorization as the root concept (partial compatibility); or none of their descendants are associated with the same two SGs as their ancestor (no compatibility).

Full compatibility. Of the 132 root concepts studied, 16 (12.1%) exhibit full compliance with inheritance principles. All their descendants are associated with exactly the same pair of SGs as the root concept itself. One such example is the root concept *Abuse of disabled person* (C0413337), associated, along with all of its descendants, with the SGs Activities & Behaviors and Disorders.

Partial compatibility. For 63 (47.7%) of the 132 root concepts studied, some of the descendants exhibit only part of the SG categorization of the root concept, i.e., are associated with only one of the SGs of the root concept (and no other SG). For example, *Periodicals* (C0031082) is associated with the SGs Concepts & Ideas and Objects. Out of its 17 descendants, four exhibit the same SGs, while 13 are associated with either Concepts & Ideas or Objects.

No compatibility. For 53 (40.2%) of the 132 root concepts studied, some of the descendants exhibit complete semantic incompatibility with the root concept (i.e., SG not shared with the root concept). For instance, *Psychomotor Performance* (C0033923), associated with the SGs Activities & Behaviors and Physiology, has:

- Three descendants associated with the same SGs;
- Fourteen descendants associated with either Activities & Behaviors or Physiology;
- Nine descendants associated with neither SG, but rather with a different SG such as *Ambidexterity* (C0002416), associated with the SG Disorders.

[6] 7 MSG root concepts were ignored, because they had more than 20,000 descendants, making the analysis unreliable.

Overall, of the 9,025 descendants of the 132 root MSG concepts studied, 8.1% fully inherit the SGs of the root MSG concept (i.e., both SGs), 64.3% partially inherit the SGs of the root MSG concept (i.e., either SG), and 27.6% inherit none of the SGs of the root MSG concept (i.e., neither SG).

4.4 Qualitative auditing of concept categorization and polysemy

We analyzed the MSG concepts in order to determine if they correspond to polysemous concepts or rather to miscategorized concepts. Of the 1,208 MSG concepts, only 91 (7.5%) were inappropriately categorized. For example, the concept *Wheat preparation* (C1095911) is associated with the SGs Chemicals & Drugs and Living Beings because it is categorized by both Pharmacological Substance and Plant. The latter categorization is not correct since *Wheat preparation* is not a kind of plant. Therefore, this concept should only be associated with the SG Chemicals & Drugs. A detailed analysis of categorization errors and illegal SG combinations is provided in section 5.1.3.

Overall, most of the MSG concepts whose multiple categorization was deemed correct are polysemous *by convention* (94.5%). For example, the concept *Computer Systems* (C0009612) is polysemous because it represents both a physical entity (categorized by Manufactured Object from the SG Objects) composed of computers and hardware, and an abstract entity (categorized by Intellectual Product from the SG Concepts & Ideas) with which data can be exchanged and processed. The polysemy of *Computer Systems* is thus characterized as polysemy *by convention* because both notions, physical and abstract, are present in the definitions provided for this concept in various sources. For example, in MeSH: “systems composed of a computer or computers, peripheral equipment, such as disks, printers, and terminals, and telecommunications capabilities” and in the National Cancer Institute (NCI) Thesaurus: “a set of hardware and software which processes data in a meaningful way”.

Only 61 MSG concepts (5.5%) are polysemous as a result from the UMLS integration process. Note that most of these MSG concepts come from multiple sources. They are actually associated with distinct SGs because they represent the different senses defined in the sources they come from (polysemy *by integration*). Like *Computer Systems*, the concept *Sculpture* (C0036473) is categorized by the STs Manufactured Object and Intellectual Product and comes from several source vocabularies. However, while *Computer Systems* is polysemous in several sources, the polysemy of *Sculpture* comes from the *integration* of distinct senses from several sources into a single concept in the UMLS Metathesaurus. Namely, *Sculpture* represents both the object in the Psychology Indexing Terms vocabulary (through a relation *child of Products of the Arts* (C0220786)) and the art realization in the Alcohol and Other Drug Thesaurus (through a relation *child of Visual arts* (C0150824)).

We also audited the categorization of MSG concepts in comparison to that of their parents. We thus restricted the analysis to those MSG concepts having at least one parent, i.e. 1,038 concepts. We counted MSG concepts whose parent categorization is consistent, inconsistent, or incomplete (Table 6). For half of MSG concepts, we found the parent categorization to be consistent. For example, the direct ancestor of the MSG concept *Perceptual Motor Performance* (C0030978) is *Psychomotor Performance* (C0033923) and they are both associated with the SGs Activities and Physiology, which was deemed consistent. For 37.7% of MSG concepts, the categorization of their parents was inconsistent. For instance, one of the parents of the MSG concept *Sexual Harassment* (C0162790) is *Criminal Offenses* (C0935509),

categorized by the ST Classification. This ST belongs to the SG Concepts & Ideas and appears to be inappropriate for the categorization of *Criminal Offenses*. For less than 13% of MSG concepts, the parent categorization was deemed incomplete. As an illustration, *RNA Sequence* (C0162327), which is associated with both Chemicals & Drugs and Genes & Molecular Sequences, is the child (among others) of *Base Sequence* (C0004793), which is only associated with the SG Genes & Molecular Sequences. Actually, the parent concept *Base Sequence* should also be categorized (e.g., like *RNA Sequence*) by Nucleic Acid, Nucleoside, or Nucleotide and should thus be associated with the SG Chemicals & Drugs.

5 Discussion

5.1 Discussion of findings

5.1.1 Characteristics of MSG concepts

One of the principles underlying the creation of the SGs is **exclusivity** [10], meaning that each Metathesaurus concept is expected to be classified into one and only one SG. Therefore, by design, the number of concepts associated with more than one SG is limited. When the SGs were introduced in 2001, there were 4,913 MSG concepts in the Metathesaurus [8]. While much larger, the 2008AA version of the UMLS comprises only 1,208 MSG concepts. In fact, the proportion of MSG concepts in the Metathesaurus actually decreased about 10 fold between 2001 (0.7%) and 2008 (0.08%).

While 10,853 UMLS concepts (0.7%) are categorized by more than two STs, no concepts are associated with more than two SGs. The wide variety of combinations of STs observed is in contrast to the limited number of combinations of SGs. Additionally, this study revealed that some pairs of categories are more frequent than others, while most combinations are never observed. A total of 30 combinations of two SGs are actually encountered in the UMLS Metathesaurus, i.e., 28.5% of the 105 possible pairs of SGs. Moreover, fifteen pairs of SGs represent 95% of all MSG concepts.

75% of all MSG concepts are associated with only four pairs of SGs (Table 3). Frequent combinations of SGs reflect patterns in how the different senses of polysemous concepts derive from each other and how they are systematically related, e.g., through metonymy. For example, *Hospitals* and other medical institutions can be understood as both physical objects and organizations, *Books* and other publications are both physical objects and intellectual products, and *Tests* are both conceptual objects and procedures.

5.1.2 Cluster interpretation

Clusters represent sets of semantically homogeneous concepts with some kind of hierarchical organization. Large clusters denote major kinds of polysemous concepts. For example, *Hospitals* and other medical institutions are grouped into a coherent tree structure and consistently categorized. The corresponding pair of SGs, Objects-Organizations, corresponds to the largest cluster (424 MSG concepts). Conversely, due to a lack of hierarchical organization in terminologies such as LOINC [21], a vast majority of clusters for the Concepts & Ideas-Physiology combination are singleton clusters. Examples of such clusters include *Cigarettes smoked, total (pack/yr)* (C0489470) and *RR interval* (C0489636), understood as temporal, qualitative, or quantitative concepts, as well as physiological items.

Small clusters may correspond to partially consistent hierarchies, where some concepts exhibit a given SG while others do not. For example, as shown in Figure 6, *Violence* (C0042693) is double-typed as Social Behavior and Mental or Behavioral Dysfunction; as such, it falls under two SGs: Activities & Behaviors and Disorders. Among its first level descendants, some are classified in a consistent manner, e.g., *AODR Violence* (C0814667), which is both an activity and a disorder and is thus categorized like its parent concept. In contrast, other descendants only represent activities, such as *Family violence* (C0206072), categorized by Individual Behavior and Social Behavior, thus associated only with the SG Activities & Behaviors. This example illustrates a situation where all the concepts having a common parent share one SG of this parent concept (e.g., social behaviors), but only some of them exhibit the other SG of this parent concept (e.g., mental dysfunctions).

5.1.3 Categorization errors

20 combinations of SGs present cases of miscategorization. Six of them exhibit a 100% rate of miscategorization (Chemicals & Drugs-Concepts & Ideas, Chemicals & Drugs-Occupations, Chemicals & Drugs-Physiology, Concepts & Ideas-Disorders, Disorders-Procedures, Occupations-Procedures). These intersections correspond to small sets of concepts, with up to four concepts. These results confirm the observation from Gu et al. that small sets of multiply-categorized concepts tend to correspond to miscategorization [17].

Among the 14 remaining SG pairs exhibiting categorization errors, the Anatomy-Objects SG combination has a 81.8% rate of miscategorization. Indeed, out of 11 concepts, nine are subcategories of *Blood* (C0005767) and should only be associated with the SG Anatomy. Other MSG concepts represent observable entities, such as *Ventricular End-Systolic Volume* (C0080308), leading to 60% of miscategorized concepts for the Phenomena-Physiology combination, and 40% for Phenomena-Procedures. Another set of miscategorized concepts corresponds to plant extracts that should not be associated with Living Beings, for example, *Sandalwood (substance)* (C1706570) and *Angelica Sinensis Root Extract* (C1879704). The percentage of erroneous categorization for combinations of SGs of more than 150 concepts is low (less than 10%). For the combination that has the highest number of concepts, Objects-Organizations, the rate of miscategorization is only 1.7%. In this group, only few concepts are not polysemous: some of them denote organizations that should not have been categorized also as objects, e.g., *Managed Care Organizations; Preferred Provider Organization* (C1551307), as well as equipments that should not have been categorized also as organizations, e.g., *Hearing Aid Equipment* (C1552500).

Out of the 91 categorization errors, 39 involve Concepts & Ideas. For example, *Body Constitution* (C0005886) and *Psychiatric consultation* (C1548378) should not be categorized as Concepts & Ideas. This observation suggests that concepts categorized as Concepts & Ideas would likely benefit from further auditing.

Finally, no real consensus was achieved for three concepts, i.e., *Voltammetry* (C0683134), *Turbidimetry* (C0041394), and *Scintillation Counting* (C0036406). In these cases, it was not clear whether the quantitative aspect of these procedures actually warranted the ST categorization with Quantitative Concept in addition to Laboratory Procedure.

5.2 Polysemy and clinical utility

Among the principles underlying the creation of the SGs is **utility**, defined as "the groups must be useful for some purpose". In fact, in addition to the necessary elements of ontological rigor (semantic validity, completeness, and exclusivity), the creators of the SGs selected practical principles (utility, parsimony, and naturalness) as guidelines for grouping STs. As a consequence of this trade-off, SGs typically cut across the top ontological distinctions between entity and event [22], grouping together, for example, Anatomical Abnormality (from the Entity tree) and Disease or Syndrome (from the Event tree), both clustered into the SG Disorders. However, some MSG concepts represent both entities and events, such as *Women's rights movement* (C0683632) which represents a social group (SG Living Beings), as well as an activity (SG Activites & Behaviors). Similarly, the distinction between roles and sortal types [23] is purposely ignored in the constitution of the SGs, resulting in the grouping of Finding (role) and Disease or Syndrome (sortal type) into the SG Disorders. Broad groupings such as the SGs make practical sense. For example, the Metathesaurus concepts from the SG Disorders all represent kinds of entities that, in spite of their distinct ontological nature, can all be somewhat observed, diagnosed or treated [10]. However, not all SGs accommodate the sortal types along with their roles, as illustrated by the concept *Butter* (C0006494) categorized by both Lipid (SG Chemicals & Drugs) and Food (SG Objects).

The creation of the UMLS Metathesaurus can be thought of as an exercise in semantic normalization from the perspective of clinical utility, i.e., usefulness for physicians and health professionals in clinical information systems. In other words, semantic normalization in the UMLS is guided not solely by ontological principles [24,25], but also by purpose. By imposing a "concept-oriented" view on terms from biomedical vocabularies, the terminology integration process of the UMLS overnormalizes some terms (sometimes resulting in polysemous Metathesaurus concepts), while denormalizing other terms (preventing polysemous concepts from being created in the Metathesaurus). Of course, although guided by clinical utility, maintaining balance in semantic normalization is art rather than science.

Cases of overnormalization are reported in [26] and reflected in the way substances and pharmaceutical products from SNOMED CT are integrated in the UMLS Metathesaurus. SNOMED CT distinguishes between the ingredient of a drug, e.g., cetirizine, as a substance (e.g., available in hydrochloride salt) and the drug itself, as an entity that can be purchased from a pharmacy (e.g., available in the form of chewable tablets). In SNOMED CT, most drug ingredients are represented with two distinct concepts, one for the substance and one for the product, associated with the corresponding disjoint top-level concepts, *Substance* and *Pharmaceutical / biologic product*, respectively. While ontologically meaningful, this distinction was deemed unnecessary from a clinical perspective in the UMLS [27]. As a consequence, the two hierarchies were collapsed into one during their integration into the Metathesaurus. The SNOMED CT concepts collapsed during this process (e.g., *cetirizine (substance)* and *cetirizine (product)*) are categorized by STs from the Substance hierarchy, not by the ST Clinical Drug. In practice, such polysemous drug concepts do not result in MSG concepts. However, the hierarchy in which such concepts participate is semantically heterogeneous, because the concept *Cetirizine* (C0055147) subsumes concepts categorized as Clinical Drug (e.g., *Cetirizine 5 MG Oral Tablet* (C0982629)), in addition to other pharmacologic substances (e.g., *Cetirizine Dihydrochloride* (C0700480)).

Conversely, the UMLS also denormalizes aggregates of concepts encountered in such biomedical vocabularies as MeSH. For example, the MeSH descriptor *Teratoma* denotes a variety of types of neoplasms "composed of a number of different types of tissue, none of which is native to the area in which it occurs". The entities listed as entry terms in MeSH are identified as distinct concepts in the UMLS and further subclassified. For example, *Cystic Teratoma* and *Mature Teratoma*, listed as entry terms for *Teratoma* in MeSH, are appropriately integrated as children of *Teratoma* (C0039538) in the UMLS. Although there is no difference in categorization among these concepts (all categorized as Neoplastic Process), it should be noted that the aggregation created in MeSH for the purpose of information retrieval is modified to fit the clinical utility requirement of the UMLS Metathesaurus.

5.3 Application to auditing concept categorization

The analysis of inheritance patterns among polysemous concepts, as well as between polysemous concepts and their ancestors and descendants in the UMLS Metathesaurus provides a framework for auditing concept categorization. In fact, because we focus on MSG concepts in this study, we can easily identify polysemous concepts. Based on the inheritance principles presented earlier, it is possible to identify potential semantic mismatches between one MSG concept and its parents, on the one hand, and its descendants, on the other.

Semantic mismatches with parent concepts. From Corollary 1, we can verify whether each MSG concept inherits all its SG categorization from its parent concepts. Two types of mismatches can be observed. First, one SG present in one of the parent concepts is not inherited by the MSG concept under investigation. For example, the MSG concept *Clinical Laboratory Information Systems* (C0008962) is associated with the SG pair Concepts & Ideas-Devices while one of its parents, *Hospital Information Systems* (C0019972), is associated with the SG pair Concepts & Ideas-Objects. This is due to the fact that, unlike its parent concept, *Clinical Laboratory Information Systems* is categorized as Medical Device rather than Manufactured Object. Second, the MSG concept exhibits SGs not inherited from its parent concepts. For instance, the concept *Technical college* (C0557810) is associated with the SG pair Objects-Organizations, while its unique parent concept, *College* (C0557806), is only associated with the SG Objects. Unlike *College*, *Technical college* is categorized as Organization, in addition to Manufactured Object. Overall, excluding the 170 MSG concepts with no parent, 373 of the 1,038 remaining MSG concepts (35.9%) are fully consistent with the SG categorization of their parent concept(s). In 375 cases (36.1%), some SG from the parent concept(s) is not inherited by the child and in 507 cases (48.8%), the child exhibits some SG that is not inherited from its parent concept(s). Of note, a double inconsistency (some SG from the parents not inherited by the child and additional SG in the child not inherited from any parent) is observed in 217 cases (20.9%). In the qualitative auditing, we further analyzed the categorization of the MSG concepts' parent(s). We showed that for 12.7% of MSG concepts, the categorization of their parent(s) is incomplete and for 37.5%, it is incorrect.

Semantic mismatches with children concepts. As shown earlier, when we compared the SG categorization of a given MSG concept to that of its descendants, we observed that only 8.1% of all descendants exhibit exactly the same SG categorization (i.e., set of SGs) as the source MSG concept. One such example is the concept *Abuse of disabled person*, presented earlier. This concept and all its descendants are associated with the SG pair Activities & Behaviors-Disorders. In almost two thirds of the

cases, a descendant concept is associated with only one of the two SGs of its ancestor, the root MSG concept used as the reference. For instance, the concept *Smoke* (C0037366), associated with the SG pair Chemicals & Drugs-Phenomena, has eight descendants, each of which is associated with either Chemicals & Drugs (e.g., *Tobacco smoke* (C0439994)) or Phenomena (e.g., *Tobacco Smoke Pollution* (C0040334)). More problematic are those cases where a given descendant exhibits SGs not present in the ancestor without inheriting any of the SGs of the ancestor. For example, the concept *Serum/plasma protein finding* (C1287377), associated with the SG pair Phenomena-Procedures, has thirty descendants, none of which are associated with either Phenomena or Procedures. One such descendant is *Increased serum protein level* (C0301678), associated with the SG Disorders.

Another example of miscategorization identified through our quantitative auditing is *Arterial blood* (C0229665). This concept is part of a cluster comprising nine MSG concepts (e.g., *Systemic arterial blood*, *Pulmonary artery blood*, etc.), all associated with the SGs Anatomy and Objects. Like their ancestor *Blood*, these concepts are all categorized as Tissue (from the SG Anatomy). However, unlike *Blood*, these concepts are also categorized as Substance (from the SG Objects). Some other descendants of *Blood* are categorized as Body Substance (from the SG Anatomy), not Substance. One possible explanation is that *Arterial Blood* and the other concepts from this cluster have been wrongly categorized as Substance (instead of Body Substance). Of note, should their categorization be reverted from Substance to Body Substance, these concepts would still be multiply categorized (as Tissue and Body Substance), yet would no longer be MSG concepts, because both STs belong to the SG Anatomy.

5.4 Limitations

As for the UMLS Metathesaurus in general, the SGs have been created from the perspective of clinical utility. For other purposes, however, other groupings have been proposed (e.g., [18]). This study would yield different results if it were based on different groupings. Therefore, to some extent, the use of polysemous, multiply-categorized concepts for auditing concept categorization is contingent upon the purpose underlying the creation of groupings of categories.

MSG concepts only represent a fraction of all polysemous concepts in the UMLS Metathesaurus. In fact, as mentioned earlier, the SGs were designed to absorb part of the polysemy encountered in biomedical concepts. A polysemous concept such as *Cleft palate*, categorized by both Congenital Abnormality and Disease or Syndrome, is not considered in this study, because these two STs belong to the same SG (Disorders).

As our study is SG-based, the number of concepts audited is much smaller than it would if performed at the ST level. We proved that it is worth studying concepts which are associated with more than one SG because some of them exhibit miscategorization in the UMLS (7.5%), while the others are truly polysemous (92.5%). On the other hand, because it is dependent on SGs rather than STs, the method proposed here is necessarily coarse-grained and favors precision over recall. Our method is therefore complementary to methods based on more fine-grained categories (e.g., STs) developed by other research groups. As already mentioned, Gu et al. and Cimino have studied frequent, legal associations of UMLS STs as a framework for auditing concept categorization [17,20]. The systematic auditing of concepts assigned to multiple STs remains a daunting task (218,536 concepts are assigned at least two STs – 181 times more than MSG concepts).

In general, auditing methods are at best semi-automatic and allow investigators to identify potentially miscategorized concepts and to focus the attention of the Metathesaurus editors on such concepts. Another limitation common to many categorization auditing methods is that only a fraction of all UMLS concepts are amenable to such auditing, namely multiply-categorized concepts. Although focusing on the 1,208 MSG concepts, our method forms the basis for auditing concept categorization for many other (at least 10,000) concepts found in the ancestors, descendants and siblings of these MSG concepts.

6 Conclusions

This study is primarily an investigation in the representation of polysemous concepts in the UMLS Metathesaurus, with focus on those 1,208 concepts associated with multiple SGs. The major categories of polysemous concepts, whose meanings are related in more or less systematic and predictable ways, include 39% of physical objects and organizations (e.g., hospitals), 16% of physical objects and intellectual products (e.g., books), and 13% of conceptual objects and procedures (e.g., tests). The analysis of inheritance patterns of SGs between one MSG concept and its parents and descendants revealed 88% of semantic mismatches (i.e., differences in SG categorization between one MSG concept and its parents and descendants). We performed a qualitative analysis of MSG concepts and show that 94.5% of MSG concepts are polysemous *by convention* and the remaining 5.5% result from the integration process realized by the UMLS.

We also showed that the analysis of such semantic discrepancies can be leveraged for auditing concept categorization. More precisely, we found that 7.5% of MSG concepts were inappropriately categorized. Such methods could be easily implemented as part of the Metathesaurus editing environment in order to assist the Metathesaurus editors in identifying discrepancies in concept categorization. This study, although limited in scope to less than 1% of all UMLS concepts, is a systematic analysis and manual review of the concept categorization of all MSG concepts in the UMLS.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

References

[1] Pustejovsky J. The Generative Lexicon. MIT Press, Cambridge, MA, 1995.

[2] Miller GA. WordNet: A Lexical Database for English. ACM Communications , 38(11), Nov. 1995

[3] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(database issue):D267-D270

[4] Buitelaar P. CoreLex: An Ontology of Systematic Polysemous Classes. In Guarino, Nicola (ed.), Formal Ontology in Information Systems, IOS Press, 1998:221-235

[5] McCray AT, Nelson SJ. The representation of meaning in the UMLS. Methods Inf Med 1995;34:193-201

[6] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med, 1993;32(4):281-291

[7] McCray AT. An upper-level ontology for the biomedical domain. Comp Funct Genomics. 2003;4(1):80-84

[8] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Medinfo 2001:10(Pt 1):216-220

[9] McCray AT, Hole WT. The scope and structure of the first version of the UMLS Semantic Network. In Proc. Fourteenth Annual SCAMC, Los Alamitos, CA, 1990;126-130

[10] Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. Journal of Biomedical Informatics, 2003;36(6):414-432

[11] Brachman RJ. What IS-A is and isn't: an analysis of taxonomic links in semantic networks. IEEE Comput 1983;16(10):30-36

[12] Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C. Relations in biomedical ontologies. Genome Biology 2005:6(5):R46.1-R46.15

[13] Cruse DA. Lexical semantics. Cambridge, UK: Cambridge University Press; 1986

[14] Sowa JF. Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley, Reading, MA, 1984

[15] Bernauer J. Subsumption principles underlying medical concept systems and their formal reconstruction. Proc Annu Symp Comput Appl Med Care 1994:140-144

[16] Gu H, Perl Y, Geller J, Halper M, Liu L, Cimino JJ. Representing the UMLS as an object-oriented database: Modeling issues and advantages. J Am Med Inform Assoc 2000; (7)1:66-80

[17] Gu H, Perl Y, Elhanan G. Auditing concept categorizations in the UMLS. Artif. Intell. Med. 2004;31(1):29-44

[18] Perl Y, Chen Z, Halper M, Geller J, Zhang L, Peng Y. The cohesive metaschema: a higher-level abstraction of the UMLS Semantic Network.J Biomed Inform. 2002 Jun;35(3):194-212

[19] Cimino JJ. Auditing the Unified Medical Language System with semantic methods. J Am Med Inform Assoc 1998;5(1):41-51

[20] Cimino JJ, Min H, Perl Y. Consistency across the hierarchies of the UMLS Semantic Network and Metathesaurus. Journal of Biomedical Informatics 2003;36(6):450-461

[21] Forrey AW, McDonald CJ, DeMoor G, Huff SM, Leavelle D, Leland D, Fiers T, Charles L, Griffin B, Stalling F, Tullis A, Hutchins K, and Baenziger J. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. Clin Chem, 1996:42(1):81-90

[22] Kumar A, Smith B. The Unified Medical Language System and the Gene Ontology: Some Critical Reflections. In KI2003: Advances in AI, A. Günter and R. Kruse and B. Neumann, 2003:135-148

[23] Guarino N, Welty C. Ontological Analysis of Taxonomic Relationships. In: Conceptual Modeling - ER 2000. LNCS, vol. 1920, Springer Berlin / Heidelberg. 2000:273-90

[24] Pisanelli DM, Gangemi A, Massimo B. Catenacci C. Coping with Medical Polysemy in the Semantic Web: the Role of Ontologies. Medinfo. 2004;11(Pt 1):416-419
[25] Spackman K, Reynoso G. Examining SNOMED from the perspective of formal ontological principles. In: Hahn J, editor. KR-MED 2004 Proceedings. Whistler, Canada: AMIA; 2004:81-87
[26] Campbell KE, Oliver D, Shortliffe EH. The Unified Medical Language System: toward a collaborative for solving terminologic problems. J Am Med Inform Assoc, 1998:5:12-16
[27] Fung KW, Hole WT, Nelson SJ, Srinivasan S, Powell T, Roth L. Integrating SNOMED CT into the UMLS: an exploration of different views of synonymy and quality of editing. J Am Med Inform Assoc. 2005;12(4):486-94

Legends

Table 1. List of the 15 SGs with their identifier and label. The number of STs per SG is also displayed

Table 2. Number of concepts associated with each SG, ordered by frequency (NB: The total is different from the total number of concepts, i.e. 1,468,958, because concepts having more than one SG are counted several times)

Table 3. Number of MSG concepts associated with each pair of SGs and cumulative percentage

Table 4. Number of clusters by pairs of SGs. Non-singleton and singleton clusters are counted separately

Table 5. Semantic characterization of polysemous concepts. For each SG pair, the predominant ST combinations are displayed. An example is given for each combination

Table 6. Analysis of the categorization of the parents of the MSG concepts

Figure 1. Representation of the concept *Medical center* in the UMLS. The CUI of this concept, the STs to which it is assigned and the SGs containing these STs are displayed

Figure 2. (a) Vertical and (b) horizontal merging of clusters

Figure 3. Categorization inheritance principle at the SG level (the ST level is hidden for simplicity):
(a) A MSG concept inherits its SG categorization from its parent, i.e. both concepts are associated with the same SGs,
(b) A MSG concept inherits part of its SG categorization from each parent

Figure 4. SGs combinations accounting for 95% of all MSG concepts. (Only pairs of SGs with a frequency of at least 10 are represented here.) The number of MSG concepts associated with each SG is indicated in parentheses in the node. The number of MSG concepts for a given pair of SGs is displayed next to the corresponding link

Figure 5. Distribution of the number of concepts per cluster

Figure 6. Cluster of eight MSG concepts associated with the Activities & Behaviors-Disorders SG combination. (NB: The concept with a dashed frame is not a MSG concept, but participates in clustering. Dashed ellipses correspond to original clusters of hierarchically-related concepts)

Table 1. List of the 15 SGs with their identifier and label. The number of STs per SG is also displayed

Identifier	Label	Number of STs
ACTI	Activities & Behaviors	9
ANAT	Anatomy	11
CHEM	Chemicals & Drugs	26
CONC	Concepts & Ideas	12
DEVI	Devices	3
DISO	Disorders	12
GENE	Genes & Molecular Sequences	5
GEOG	Geographic Areas	1
LIVB	Living Beings	23
OBJC	Objects	5
OCCU	Occupations	2
ORGA	Organizations	4
PHEN	Phenomena	6
PHYS	Physiology	9
PROC	Procedures	7
	Total: 15	**Total: 135**

Table 2. Number of concepts associated with each SG, ordered by frequency (NB: The total is different from the total number of concepts, i.e. 1,468,958, because concepts having more than one SG are counted several times)

SG Identifier	Number of concepts
CHEM	452,377
LIVB	335,179
DISO	252,284
PROC	113,701
ANAT	94,818
PHYS	83,748
DEVI	46,083
CONC	35,035
GENE	29,472
OBJC	9,537
PHEN	8,418
ACTI	4,070
ORGA	2,879
OCCU	1,451
GEOG	1,114

Table 3. Number of MSG concepts associated with each pair of SGs and cumulative percentage

Pair of SGs	Number of MSGs	Cumulative Percentage
OBJC-ORGA	472	39.0%
CONC-OBJC	194	55.1%
CONC-PROC	162	68.5%
CONC-PHYS	82	75.3%
ACTI-DISO	37	78.4%
CHEM-OBJC	29	80.8%
CONC-DEVI	28	83.1%
ACTI-CONC	26	85.3%
PHYS-PROC	25	87.3%
ANAT-CHEM	23	89.2%
ACTI-PHYS	23	91.1%
PHEN-PHYS	15	92.4%
CHEM-GENE	12	93.4%
ANAT-OBJC	11	94.3%
CHEM-LIVB	10	95.1%
CHEM-DEVI	8	95.8%
DISO-PHYS	8	96.4%
ACTI-LIVB	6	96.9%
LIVB-OBJC	5	97.4%
CHEM-PHEN	5	97.8%
OBJC-PHEN	5	98.2%
PHEN-PROC	5	98.6%
OCCU-PROC	4	98.9%
DISO-PHEN	3	99.2%
CHEM-CONC	3	99.4%
ACTI-PROC	3	99.7%
CHEM-OCCU	1	99.8%
CHEM-PHYS	1	99.8%
CONC-DISO	1	99.9%
DISO-PROC	1	100.0%
Total: 30	**Total: 1,208**	

Table 4. Number of clusters by pairs of SGs. Non-singleton and singleton clusters are counted separately

Pair of SGs	Number of non-singleton clusters	Number of singleton clusters
OBJC-ORGA	11	23
CONC-OBJC	9	28
CONC-PROC	11	30
CONC-PHYS	2	78
ACTI-DISO	4	2
CHEM-OBJC	3	11
CONC-DEVI	2	16
ACTI-CONC	3	9
PHYS-PROC	6	9
ANAT-CHEM	4	8
ACTI-PHYS	3	6
PHEN-PHYS	4	3
CHEM-GENE	2	6
ANAT-OBJC	1	2
CHEM-LIVB	2	6

Table 5. Semantic characterization of polysemous concepts. For each SG pair, the predominant ST combinations are displayed. An example is given for each combination

Pair of SGs	Predominant STs combination(s)	Example(s)
OBJC-ORGA	(Manufactured Object - Health Care Related Organization)	*HOSPITALS AND INSTITUTIONS* (C0337951)
CONC-OBJC	(Intellectual Product - Manufactured Object)	*Information Systems* (C0021428)
CONC-PROC	(Intellectual Product - Diagnostic Procedure)	*Manifest Anxiety Scale* (C0024720)
CONC-PHYS	(Quantitative Concept - Organism Attribute) (Spatial Concept - Organism Attribute)	*Cell Size* (C0162658) *Long axis* (C0522487)
ACTI-DISO	(Social Behavior - Mental or Behavioral Dysfunction)	*Abandonment of elderly person* (C0413336)
CHEM-OBJC	(Lipid - Food) (Pharmacologic Substance - Food) (Hazardous or Poisonous Substance - Manufactured Object)	*Nut Oil* (C1518477) *RICE BRAN* (C0982374) *Fertilizers* (C0015919)
CONC-DEVI	(Intellectual Product - Medical Device)	*Medical Information Systems* (C0262877)
ACTI-CONC	(Governmental or Regulatory Activity - Intellectual Product)	*Public policy on health* (C0680811)
PHYS-PROC	(Organ or Tissue Function - Diagnostic Procedure) (Genetic Function - Molecular Biology Research Technique)	*Pulmonary Diffusing Capacity* (C0034059) *Transduction, Genetic* (C0040667)
ANAT-CHEM	(Cell - Pharmacologic Substance) (Body Substance - Amino Acid, Peptide, or Protein)	*Coactivated T Cell* (C1516687) *Bone morphogenic protein* (C0450131)
ACTI-PHYS	(Individual Behavior - Mental Process)	*Satiety Response* (C0036240)
PHEN-PHYS	(Organ or Tissue Function - Laboratory or Test Result)	*Respiratory Airflow* (C0600321)
CHEM-GENE	(Nucleic Acid, Nucleoside, or Nucleotide - Gene or Genome) (Nucleic Acid, Nucleoside, or Nucleotide - Nucleotide Sequence)	*Gene Library* (C0017272) *DNA Sequence* (C0162326)
ANAT-OBJC	(Tissue - Substance)	*Blood arterial* (C0229665)
CHEM-LIVB	(Pharmacologic Substance - Virus) (Pharmacologic Substance - Plant) (Organic Chemical - Plant)	*Recombinant Vaccinia-PSA(L155)-TRICOM Vaccine* (C1515695) *Wheat preparation* (C1095911) *Sisal fiber* (C0304070)

Table 6. Analysis of the categorization of the parents of the MSG concepts

Parent categorization	Total
Consistent	517
Inconsistent	389
Incomplete	132
Total	1,038

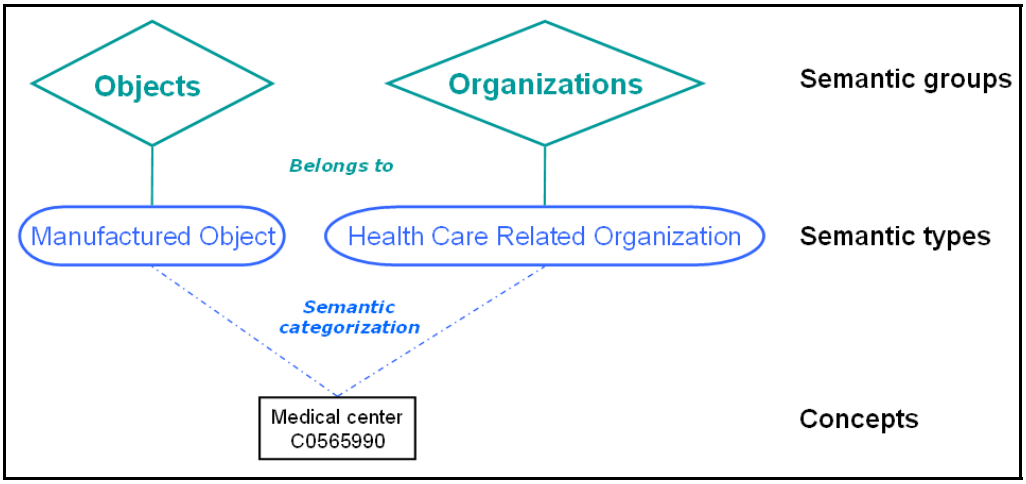

Figure 1. Representation of the concept *Medical center* in the UMLS. The CUI of this concept, the STs to which it is assigned and the SGs containing these STs are displayed

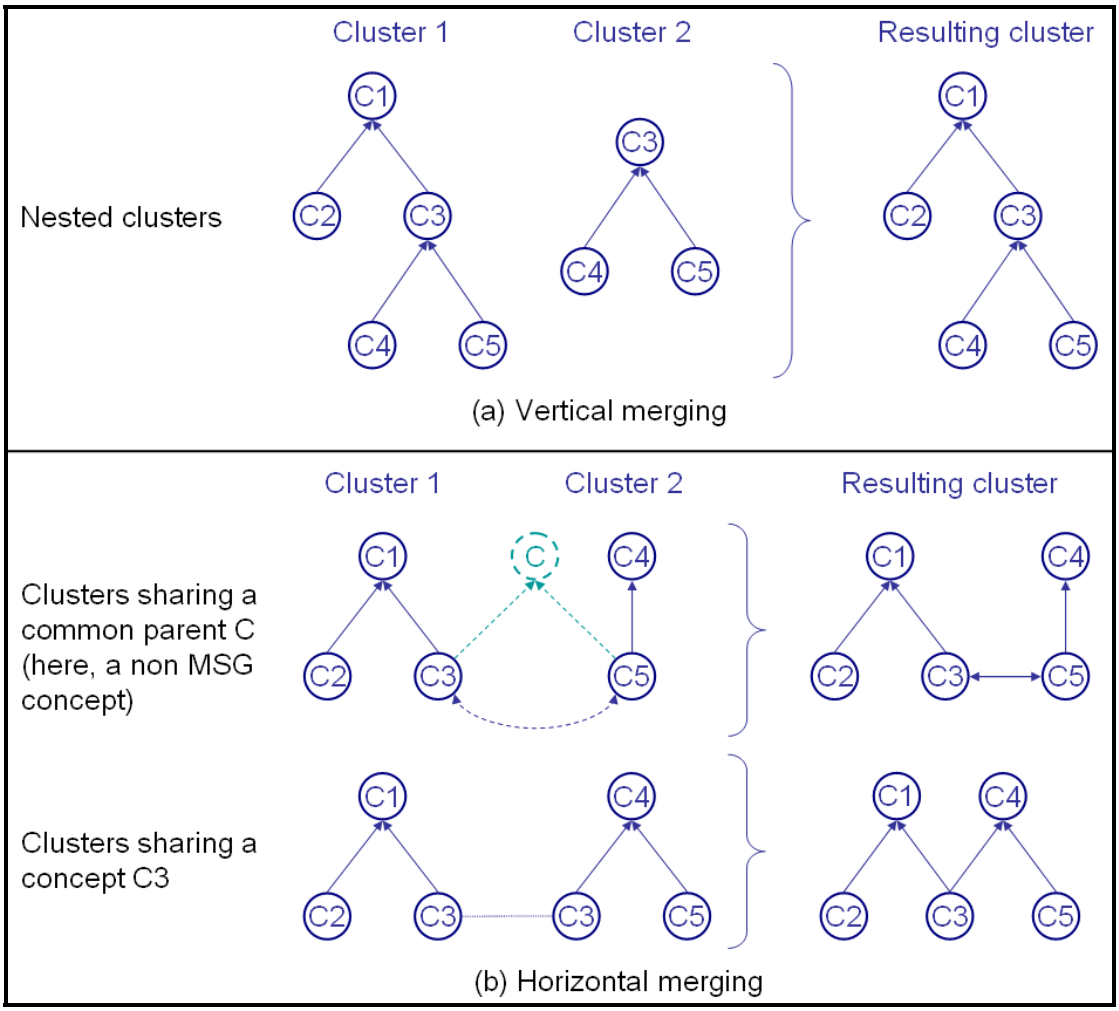

Figure 2. (a) Vertical and (b) horizontal merging of clusters

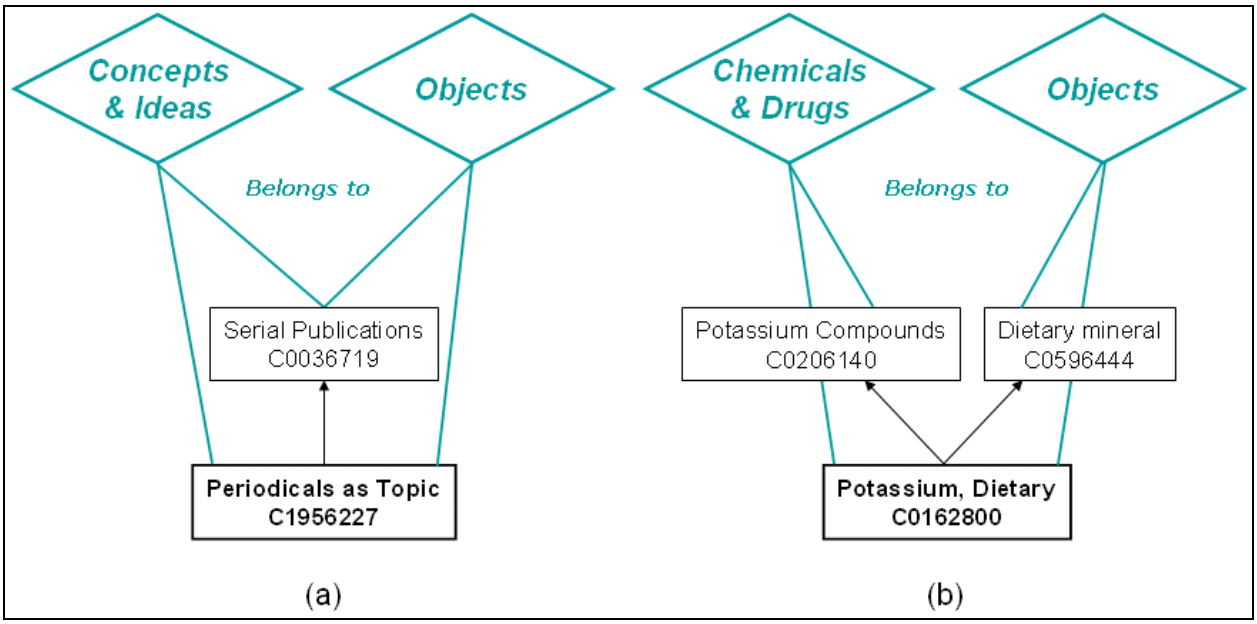

Figure 3. Categorization inheritance principle at the SG level (the ST level is hidden for simplicity):
(a) A MSG concept inherits its SG categorization from its parent, i.e. both concepts are associated with the same SGs,
(b) A MSG concept inherits part of its SG categorization from each parent

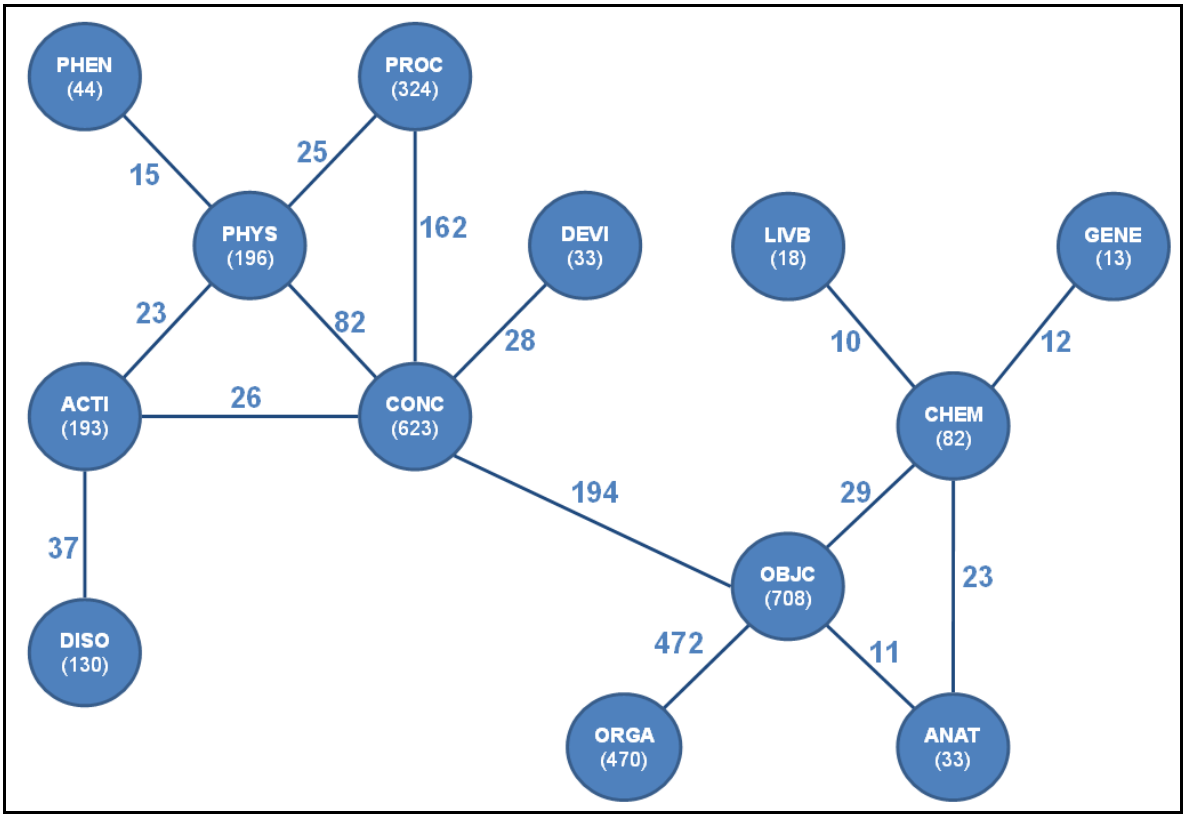

Figure 4. SGs combinations accounting for 95% of all MSG concepts. (Only pairs of SGs with a frequency of at least 10 are represented here.) The number of MSG concepts associated with each SG is indicated in parentheses in the node. The number of MSG concepts for a given pair of SGs is displayed next to the corresponding link

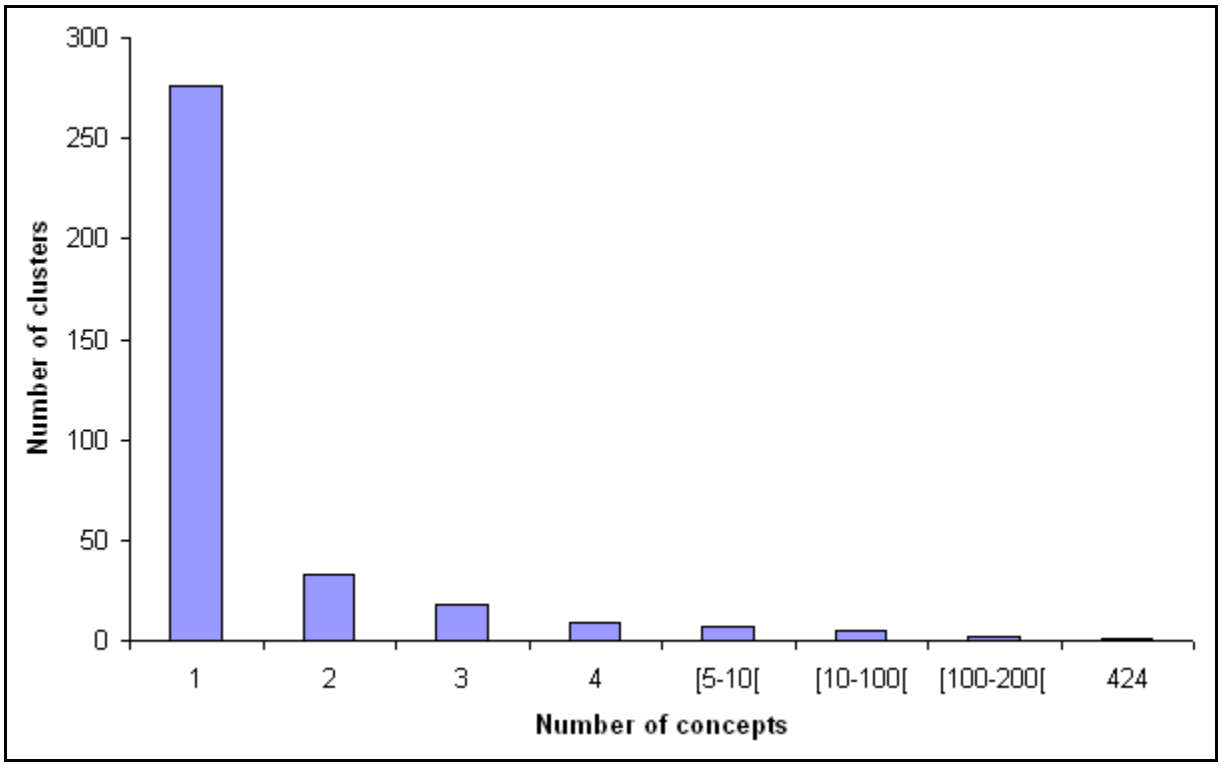

Figure 5. Distribution of the number of concepts per cluster

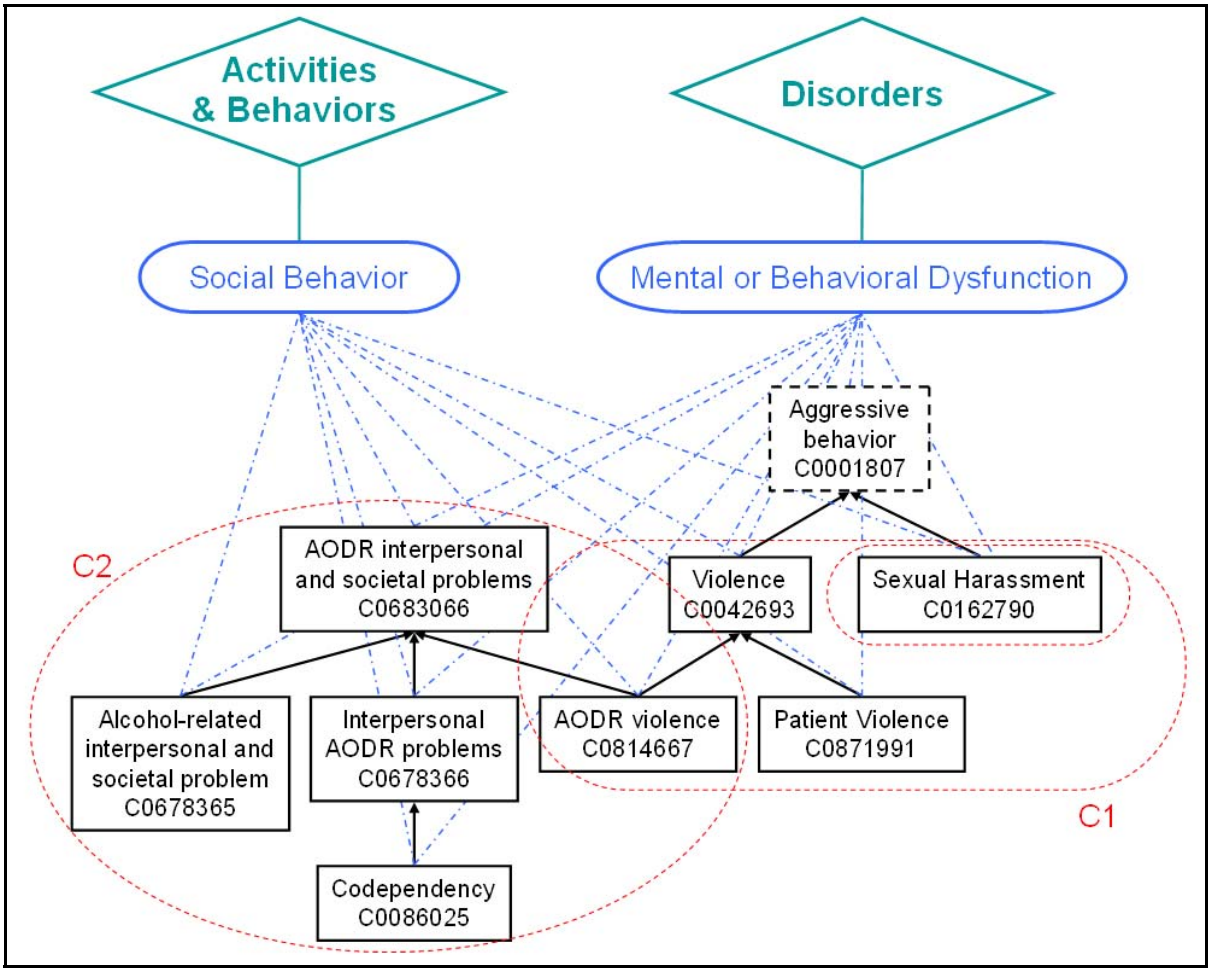

Figure 6. Cluster of eight MSG concepts associated with the Activities & Behaviors-Disorders **SG combination. (NB: The concept with a dashed frame is not a MSG concept, but participates in clustering. Dashed ellipses correspond to original clusters of hierarchically-related concepts)**